EXA3 Architecture

Daunois Jean-Claude
daunois@fr.ibm.com

John Borkenhagen
Distinguished Engineer
xSeries Server High End Architect

Agenda

- EXA Historical Overview
- X³ Basics
- X³ Design Details
- X³ Competition
- X³: Ready to Dominate

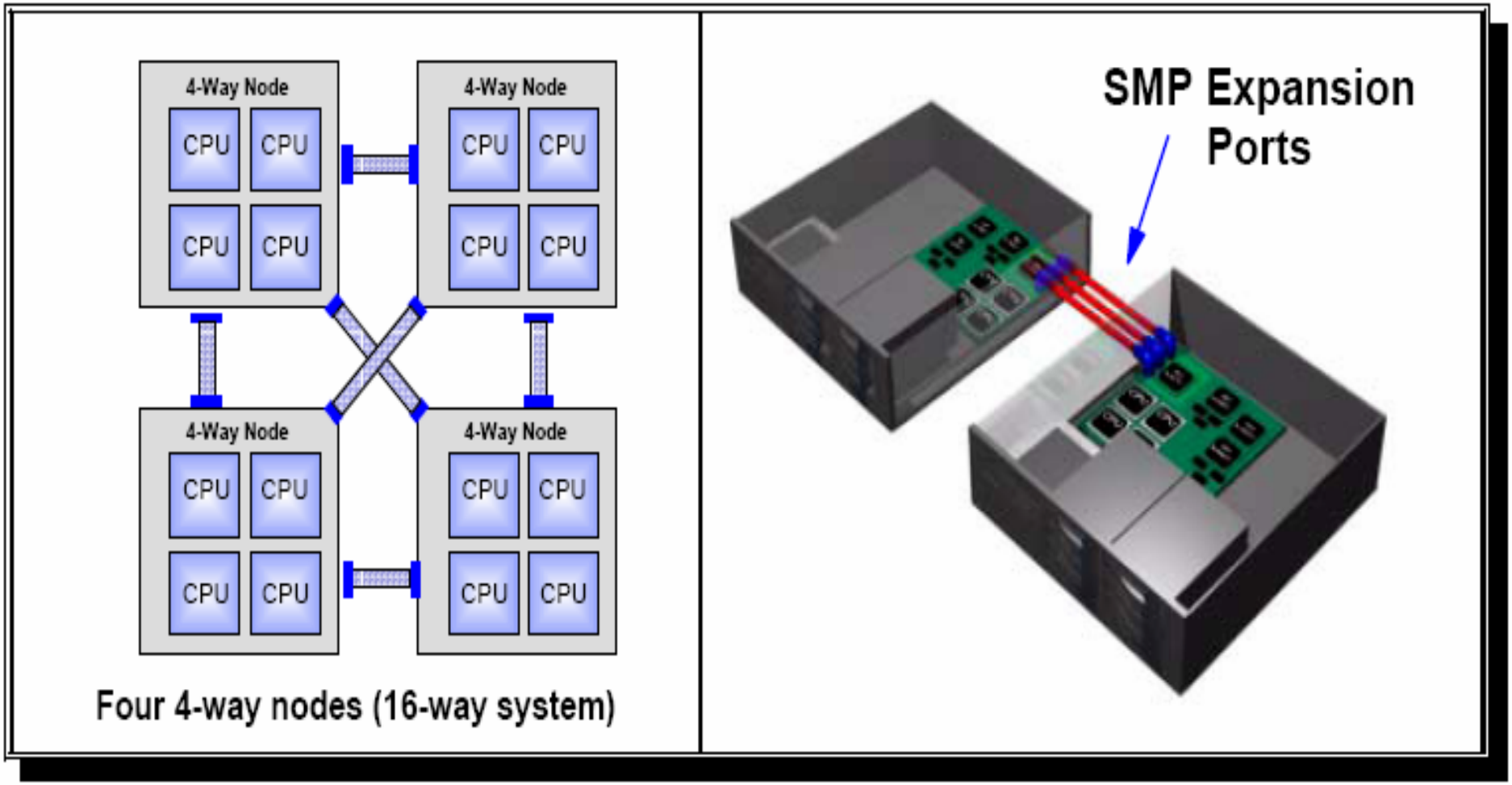
Enterprise X-Architecture™

Université des partenaires 2005

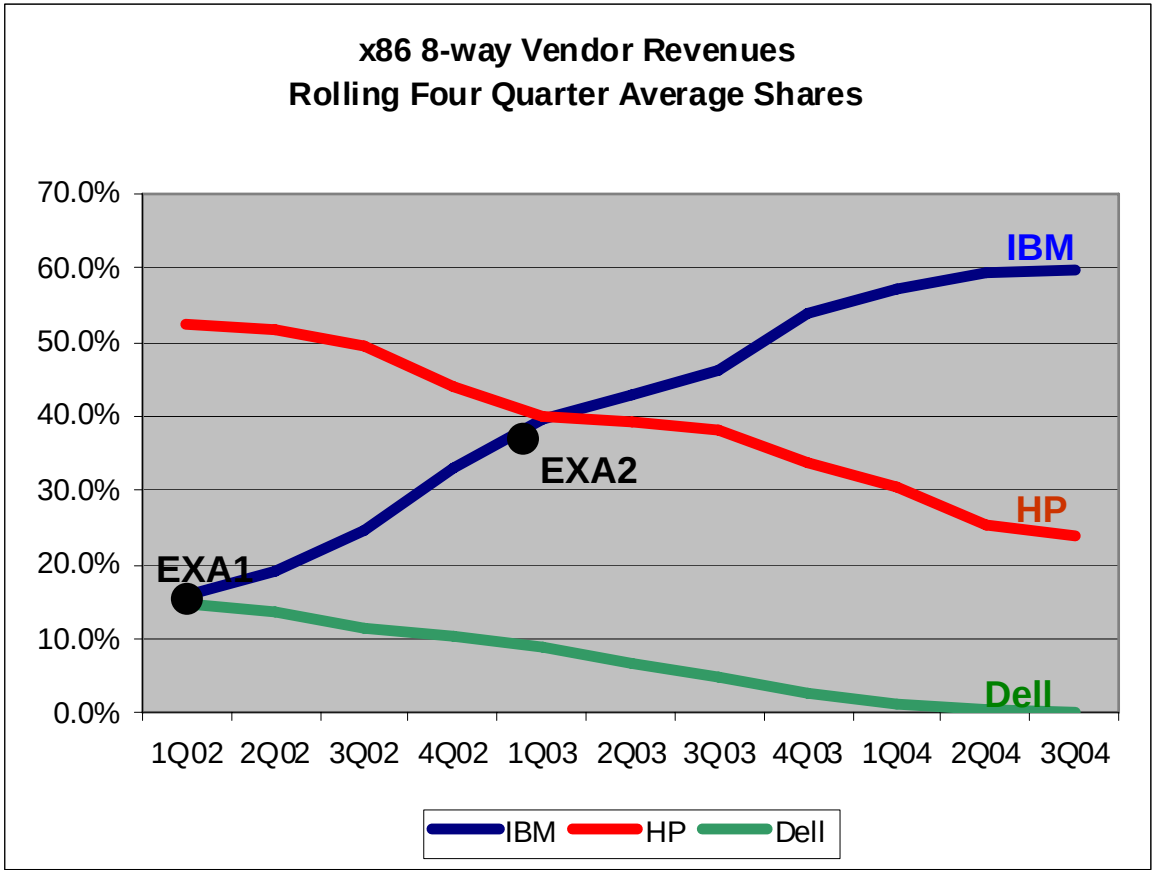
- Primary objective: Provide Intel based scalable SMP with “big iron”(1) RAS(2)
  - Commissioned in 1997
  - Resulted in innovative “pay-as-you-grow” scalable architecture
  - Incorporated “big iron” RAS features
    - Memory redundant bit steering, chip kill, mirroring
    - Redundant path run-time fail-over I/O with hot plug
    - Redundant path run-time fail-over SMP interconnect
- Unplanned EXA benefits:
  - Leadership performance starting with EXA2
  - EXA architecture enables the same or better performance with low cost small cache Intel processors (DP) compared to high cost large cache Intel processors (MP)
    - Results in superior price/performance for the customer
  - Time to market advantage
- Intel announces 64 bit extensions to IA32 in 2004
  - IA32 no longer viewed as dead end architecture, solidifies EXA future
  - xSeries no longer investing in IA64 support
  - x86-64 technology embedded into the 3/29/05 X3 GA

(1) Big iron = Mainframes (2) RAS = Reliability, Availability, Serviceability

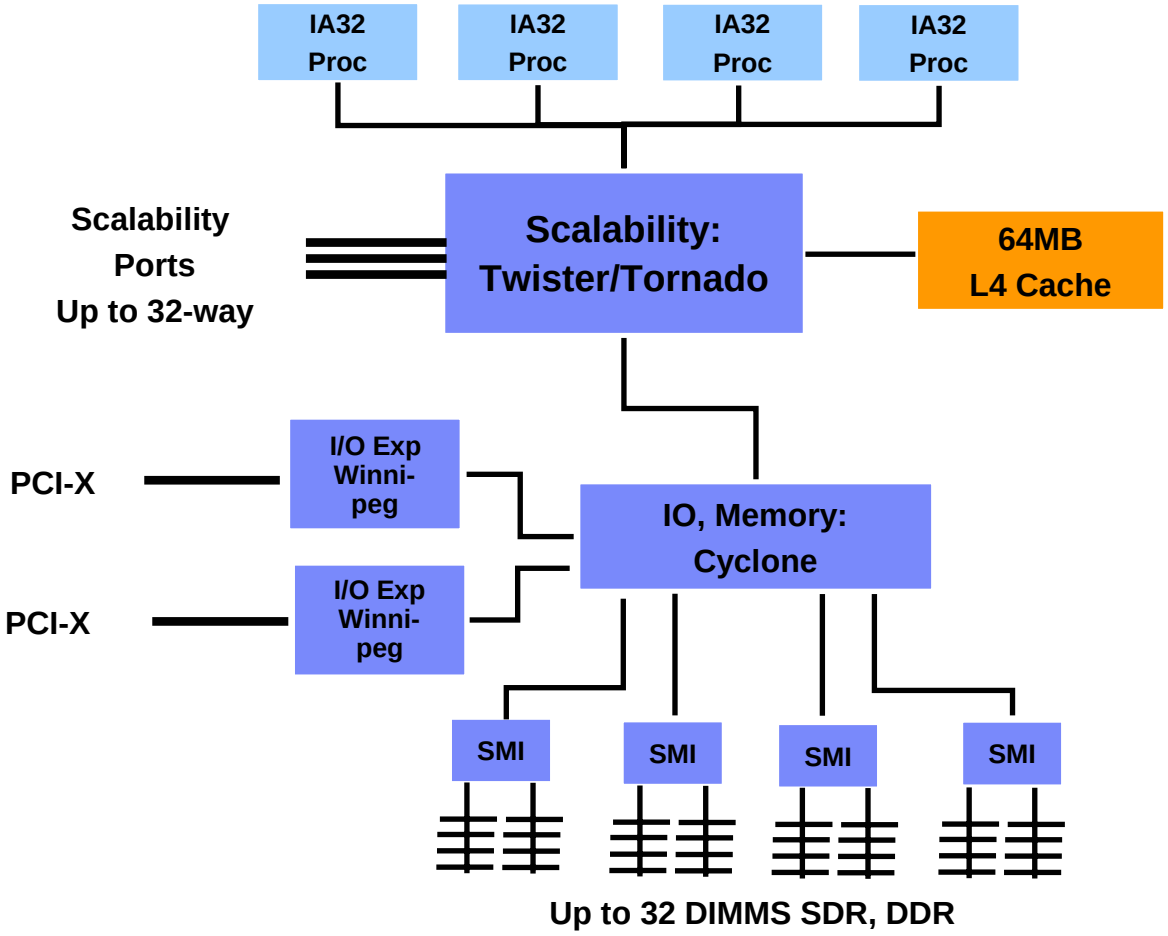
XpandOnDemand™ Scalability



- EXA has enabled IBM to Dominate the 8-way x86 Market
  - Forced Dell to exit in 2003
  - HP will exit in 2005
  - Fujitsu / NEC OEM IBM systems



EXA I: Introduction to market



- x360 4 socket IA32 piloted in December, 2001
- x440 scalable IA32 general availability April, 2002
- x450 scalable IA64 general availability May, 2003

EXA 2G

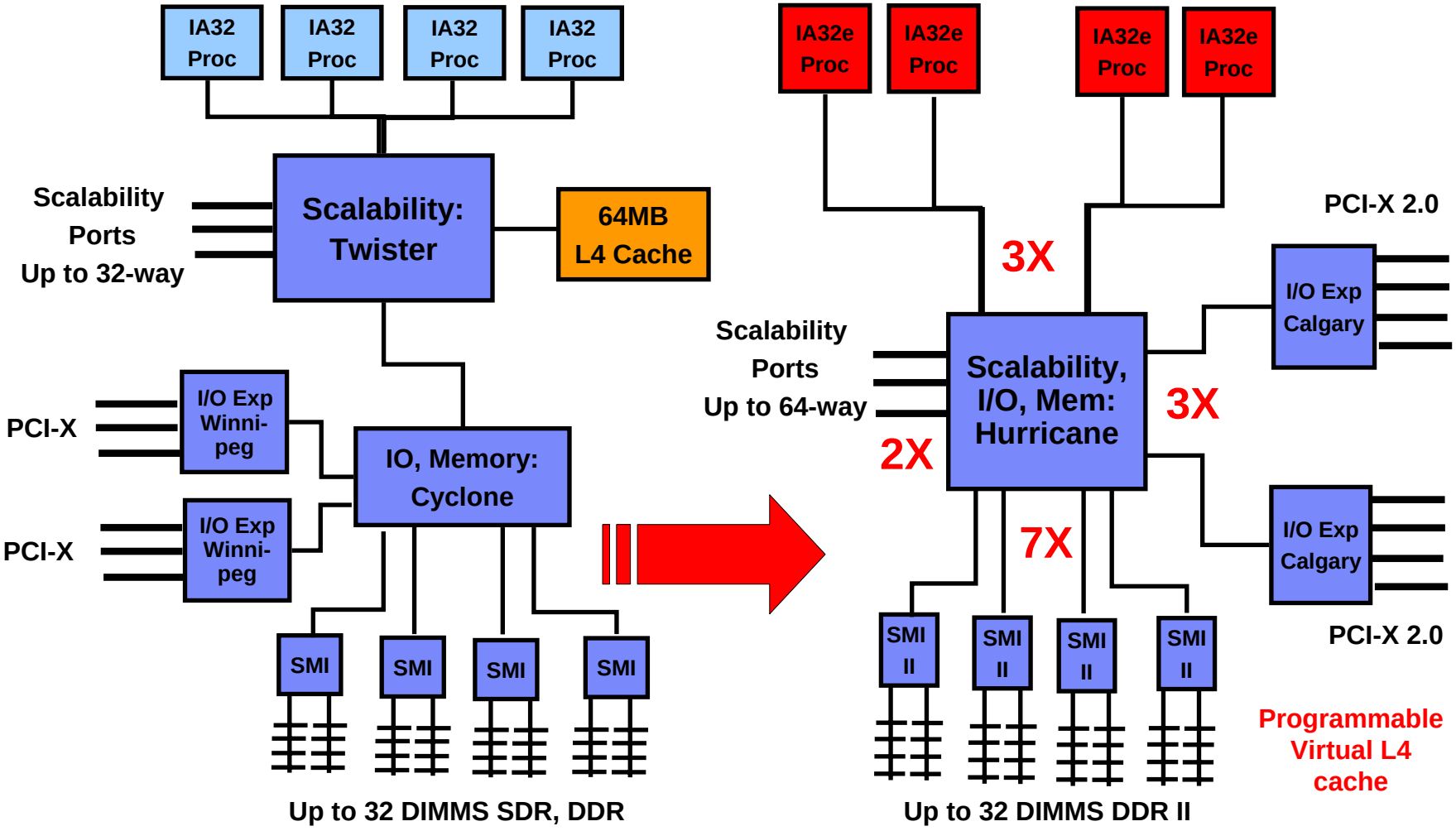
- Spin of the chips
  - Bug fixes
  - Latency reductions
  - Cost reduction: Ceramic => Laminate packaging
- New die layout and speedpath improvements
- Second set of products
  - x365 general availability December, 2002
  - x445 general availability July, 2003
  - x455 general availability November, 2003
- Dis-investment in Itanium 2
  - Dropped future Tornado funding
  - No RSA-II support for x455

X^3 - The Basics

X3: 2005 Breakthrough Performance

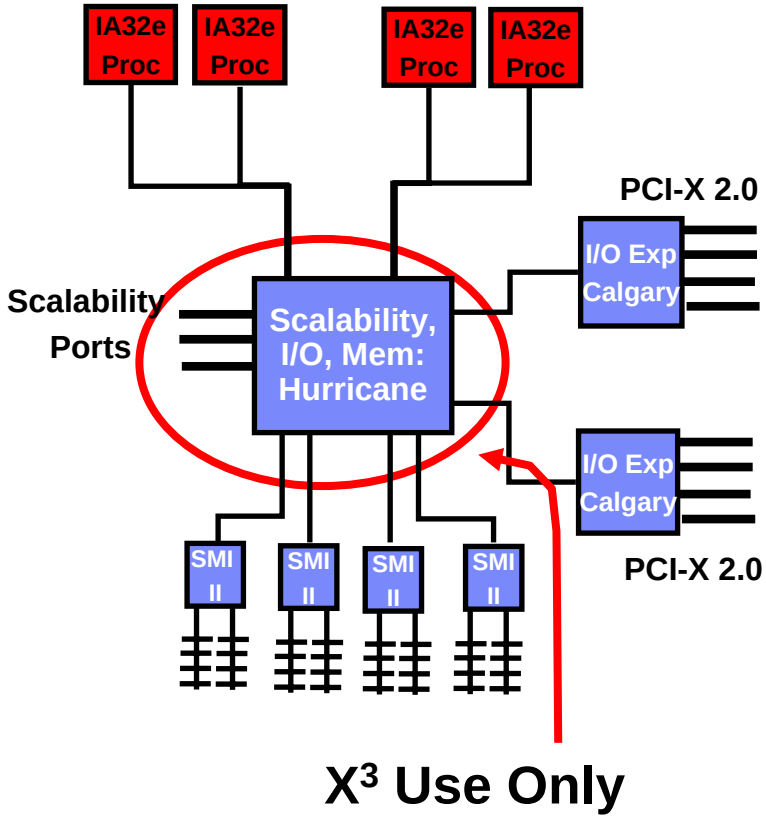
- Supports next generation Intel processors including dual-core
  - Cranford for initial release, Potomac/Cranford for scalable servers
    - Supports Intel's dual-core Paxville/Tulsa processors
  - Leverage IBM EDRAM technology
    - Incorporates snoop filter in chipset
    - Reduces latencies and traffic, especially for in-order
  - Optimized for both deferred (MP L3) and in-order (DP speed) bus operation
    - MP performance with low cost DP processors
- EM64T, DDR II and PCI-X 2.0 support added
- Scalable to 64-way SMP with Dual Core
- Significant improvements in bandwidths
  - 3X processor and I/O bandwidths
  - 2X scalability port bandwidth
  - 7X memory bandwidth
- Significantly reduced latencies for increased performance
  - 3X improvement in main memory latency
  - 3X improvement in node-to-node latency
    - SMP scaling equal to or better than a crossbar or switch structure
    - Retains all EXA modularity benefits

EXA II vs. X³ – Revolution, not Evolution



Hurricane Node Controller – Heart of X³

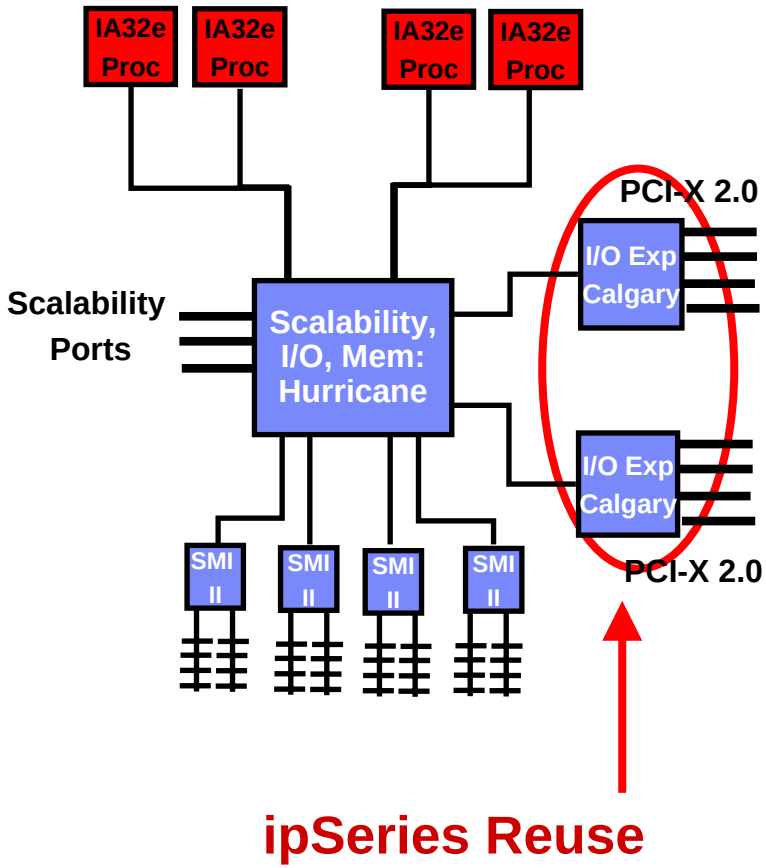
- Specifically designed for third generation X³
- Multi-site architectural team
  - Leveraged xSeries, iSeries, pSeries, and zSeries backgrounds
- Chip logic implementation by IBM E&TS
  - PowerPC processor design experience
  - Complex large SMP validation techniques
  - Maniacal focus on latency
  - "Big iron" memory and I/O controller design experience
  - "Big Iron" RAS techniques



One chip is better than two from a performance and cost perspective.

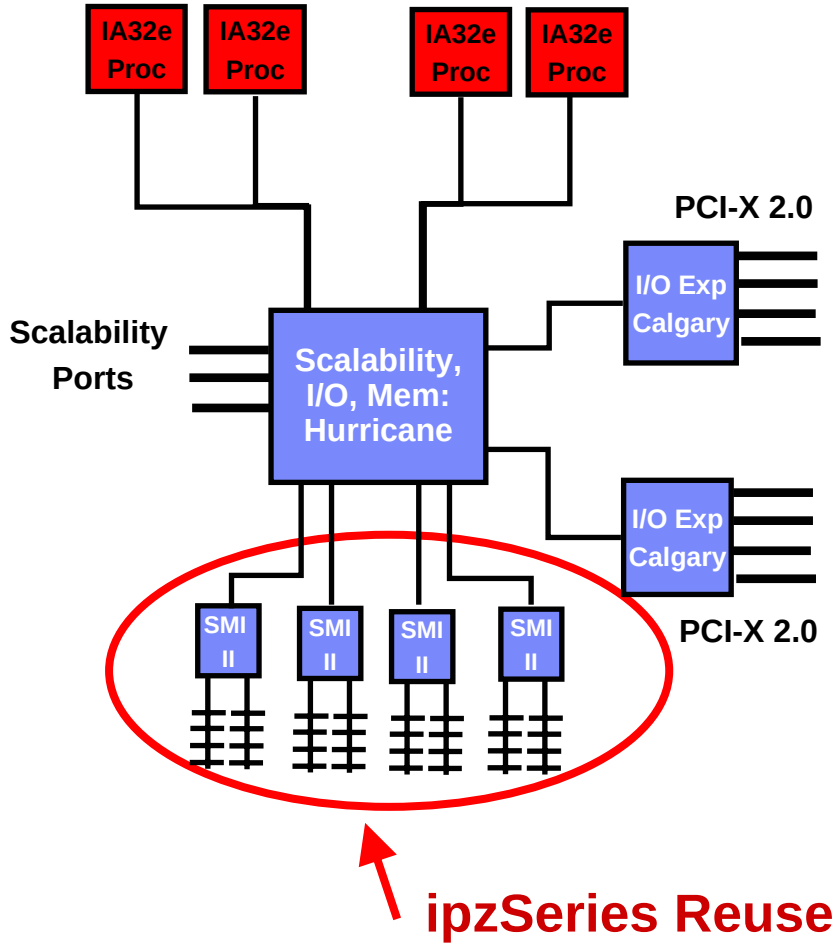
Calgary I/O Controller – ‘South bridge’

- PCI-X 2.0 I/O bridge for ipSeries
  - Minor additions for unique x86 mode of operation
- xSeries leverages “Big Iron” RAS and performance
  - Potential future enablement of ipSeries I/O virtualization logic
    - Provides RAS, security, and performance
    - Requires hypervisor enablement
    - Ideal for sharing I/O in server consolidation



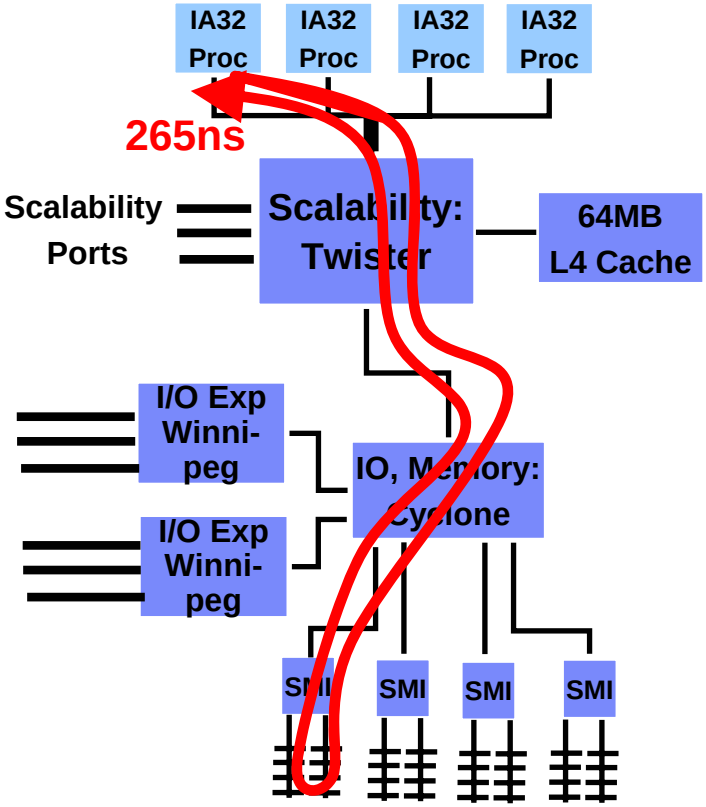
Synchronous Memory Interface (SMI) Chip

- SMI II is the memory re-drive chip for ipzSeries servers
  - Minor additions for unique xSeries mode of operation
  - xSeries leverages the “big iron” RAS and performance of SMI II

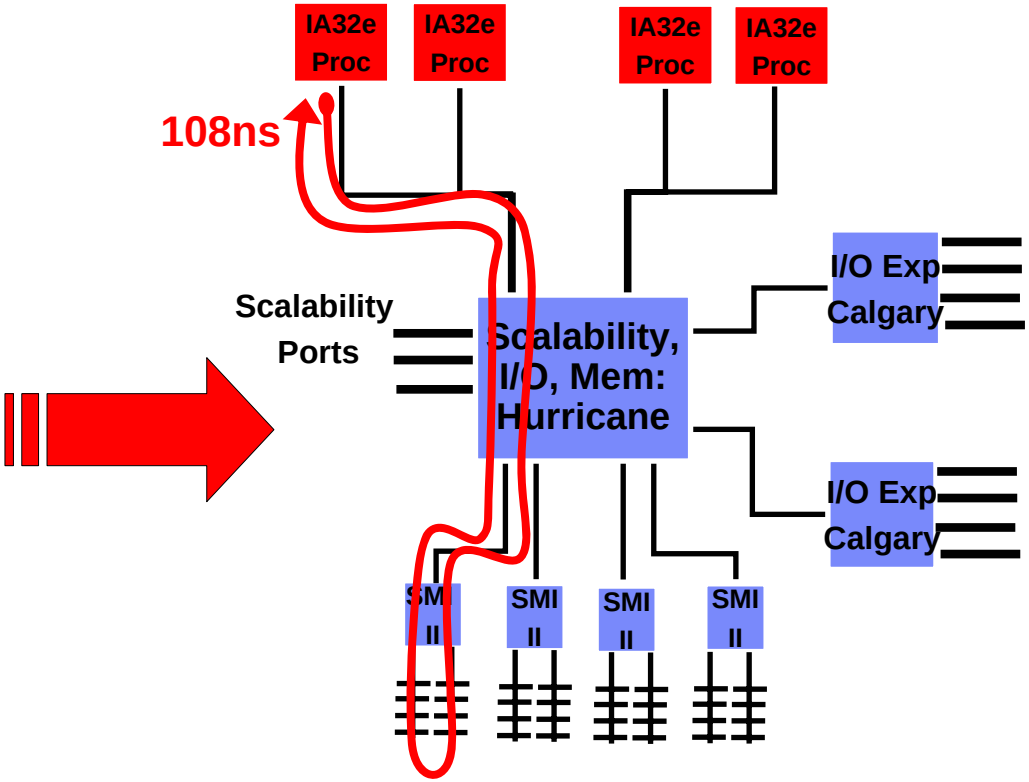


3X Main Store Latency Reduction

EXA II: 265ns



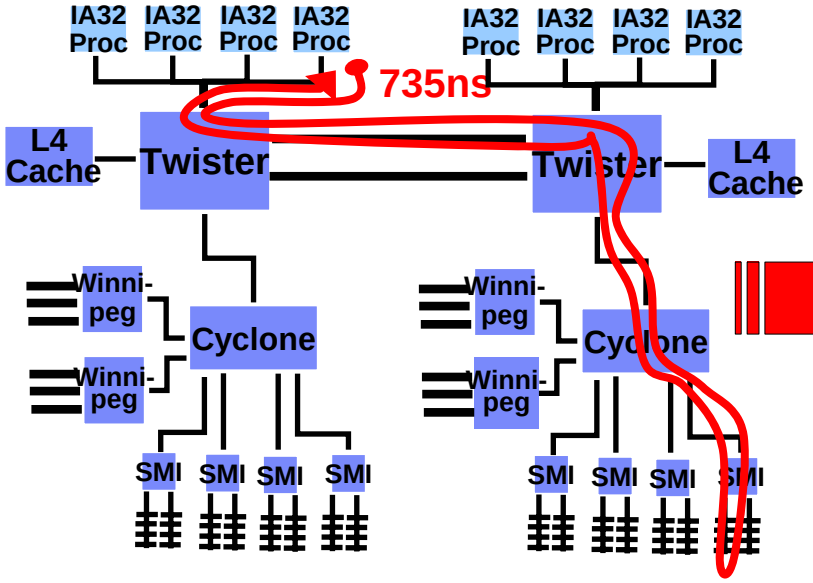
EXA III: 108ns



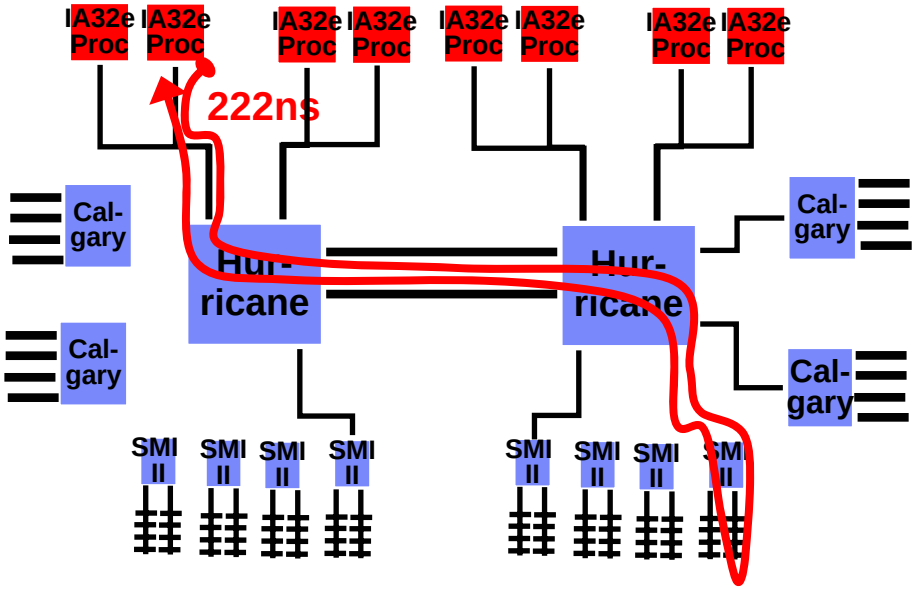
Main Store Latency Measured (1st Data)

3X Remote Latency Reduction

EXA II: 735ns

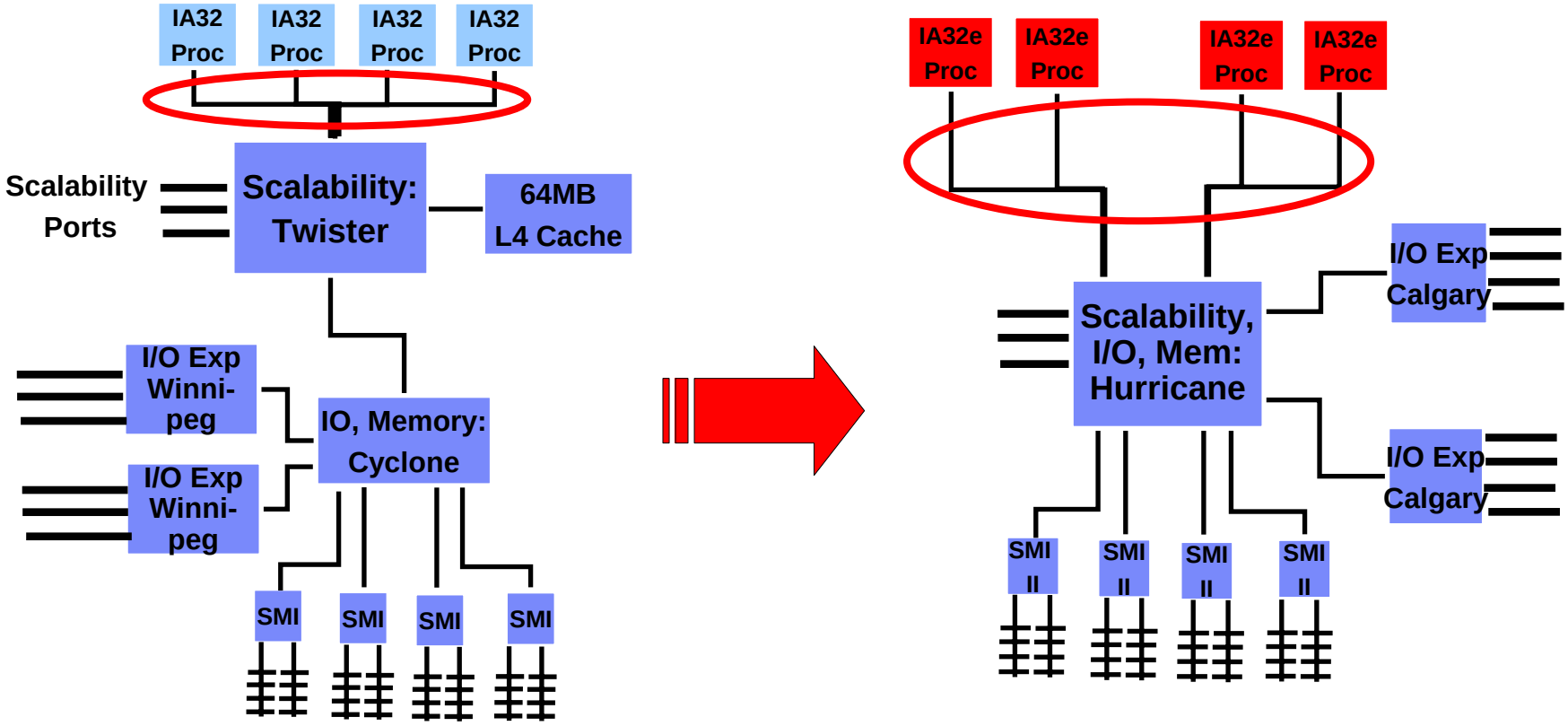


EXA III: 222ns



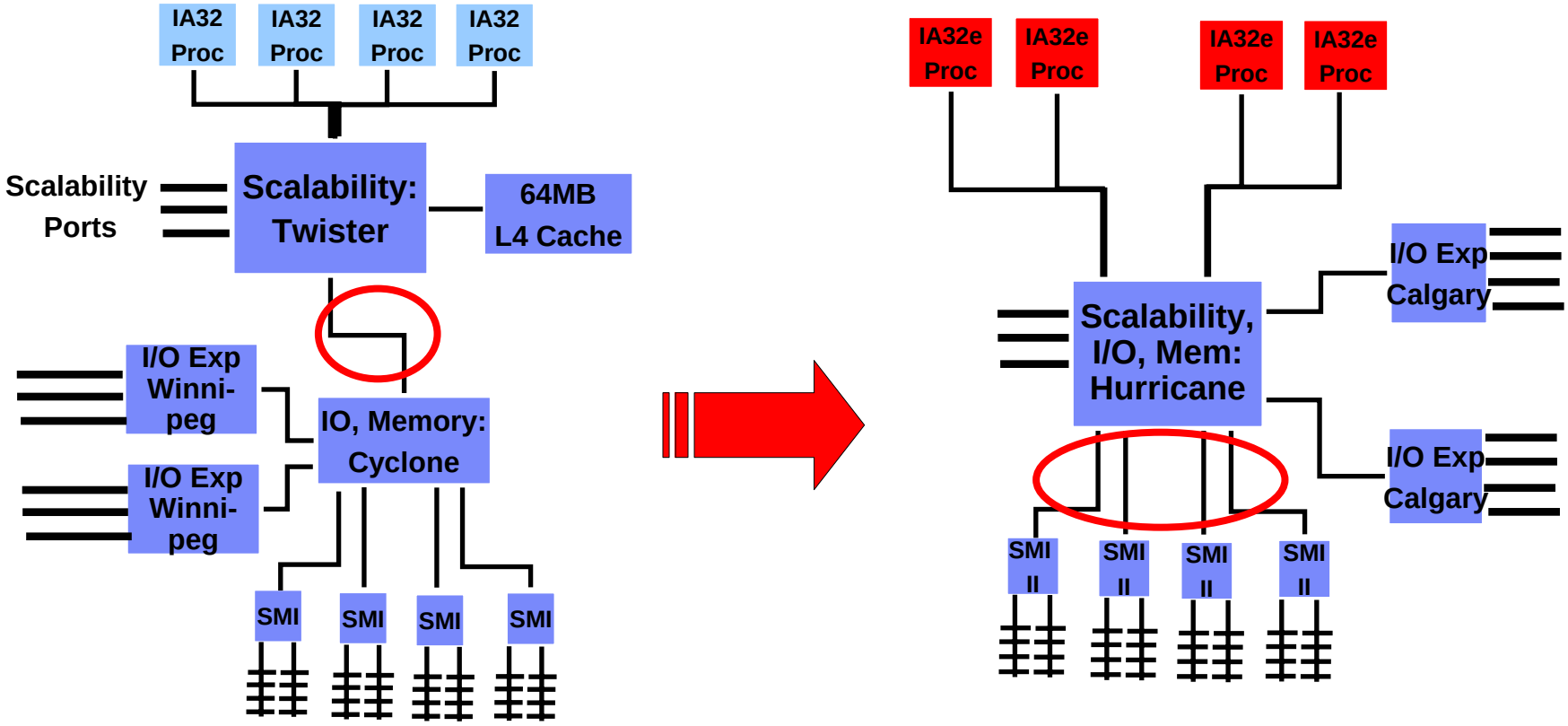
Latency Measured (1st Data)

3X Processor Bus Bandwidth



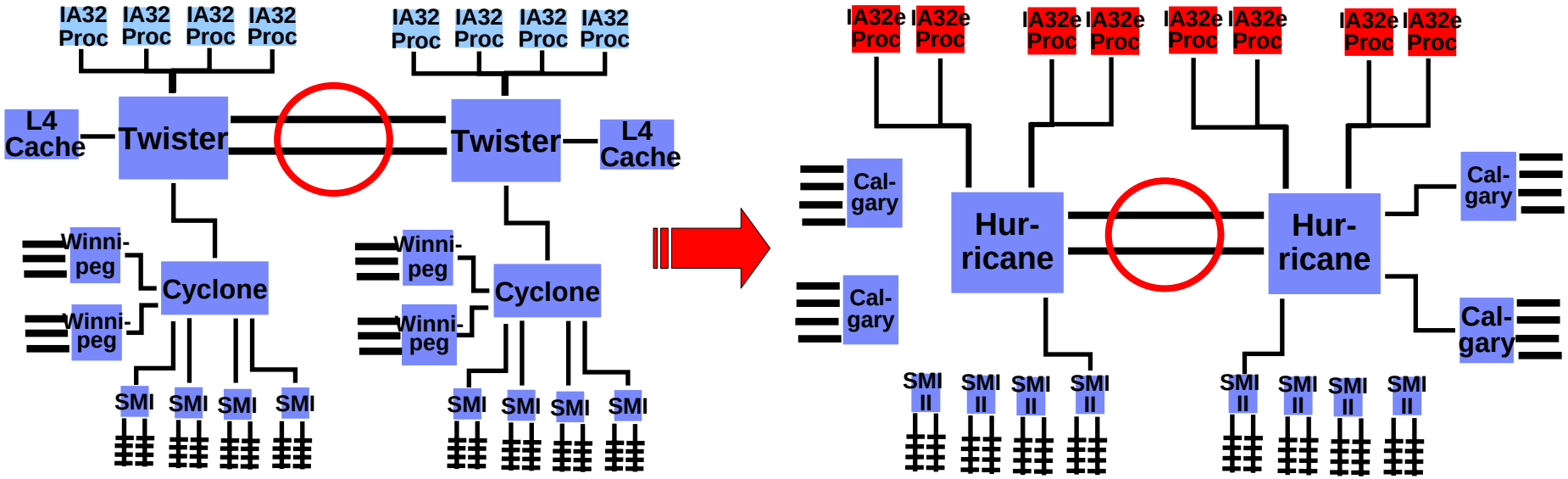
- EXA II: 8B x 400MT/s x 1Bus = **3.2GB/s**
- EXA III: 8B x 667MT/s x 2Bus = **10.6GB/s**

7X Main Store Bandwidth



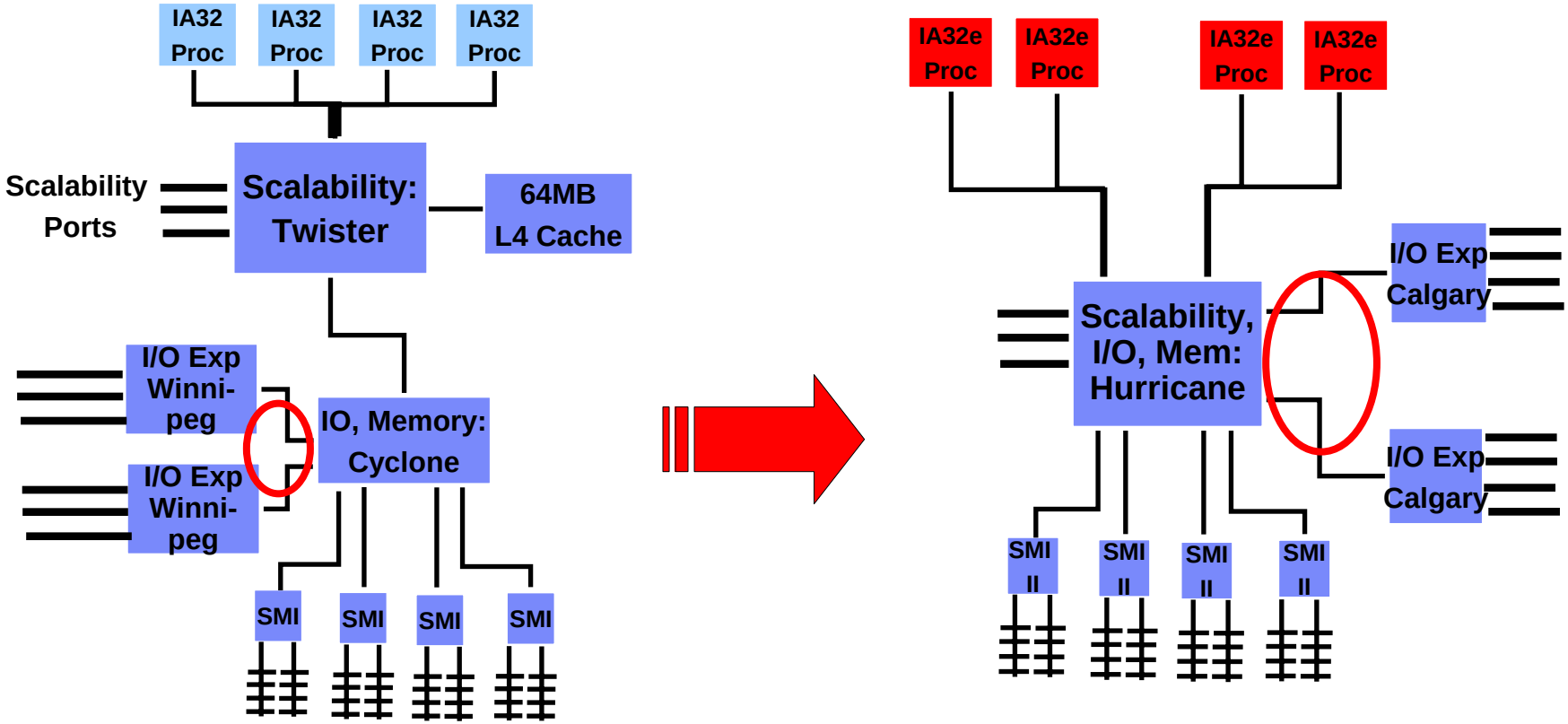
- EXA II: 8B x 400MT/s x 1Bus = **3.2GB/s**
- EXA III: 8B x 667MT/s x 4Bus = **21.3GB/s**

2X Scalability Port Bandwidth



- EXA II: 2B x 800MT/s x 2Dir = **3.2GB/s per Port**
- EXA III: 1B x 3.2GT/s x 2dir = **6.4 GB/s per port**

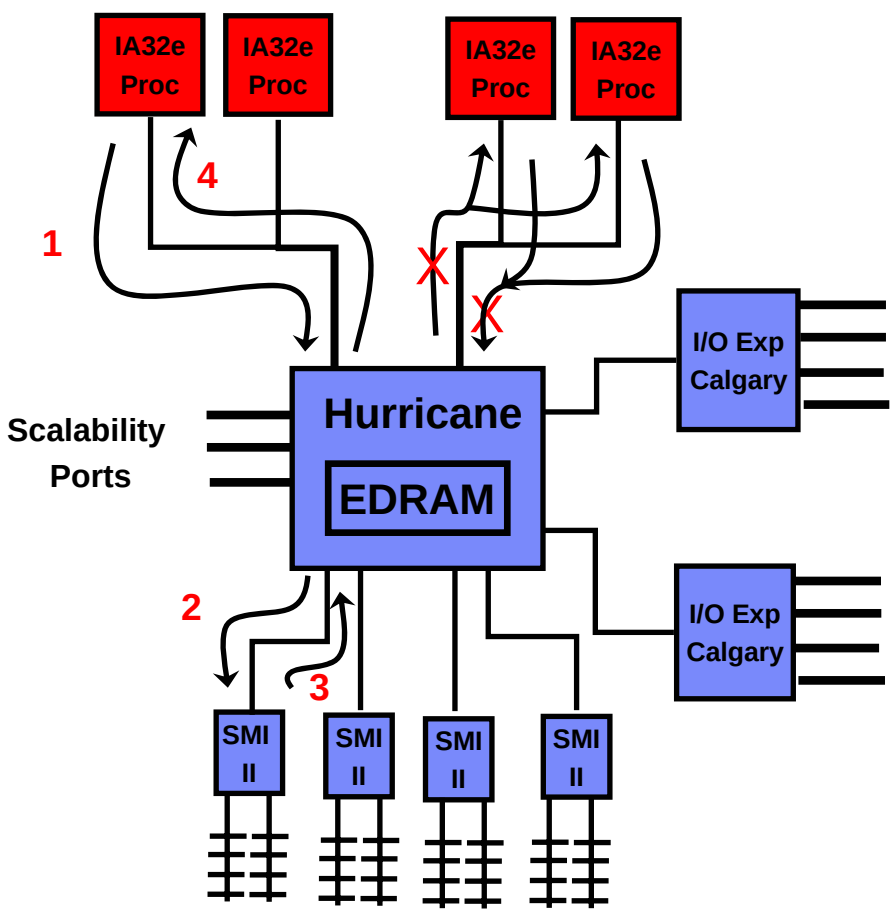
3X I/O Bandwidth

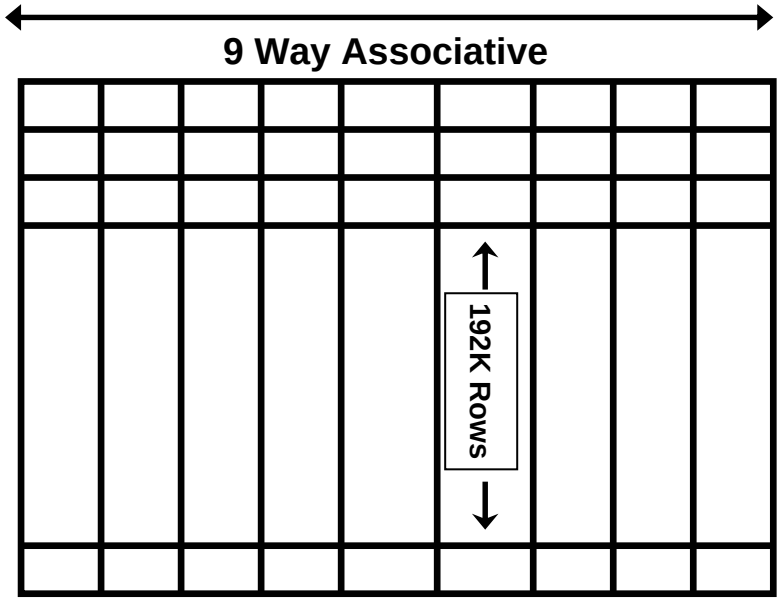


- EXA II: RIOG
  1B x 1GT/s x 2Dir x 2Bus =
  4GB/s

- EXA III: Infiniband Electrical
  1.5B x 2.5GT/s x 2 Dir X 2Bus =
  15GB/s (12GB/S after 8/10)

EDRAM Snoop Filter: 4 socket Operation

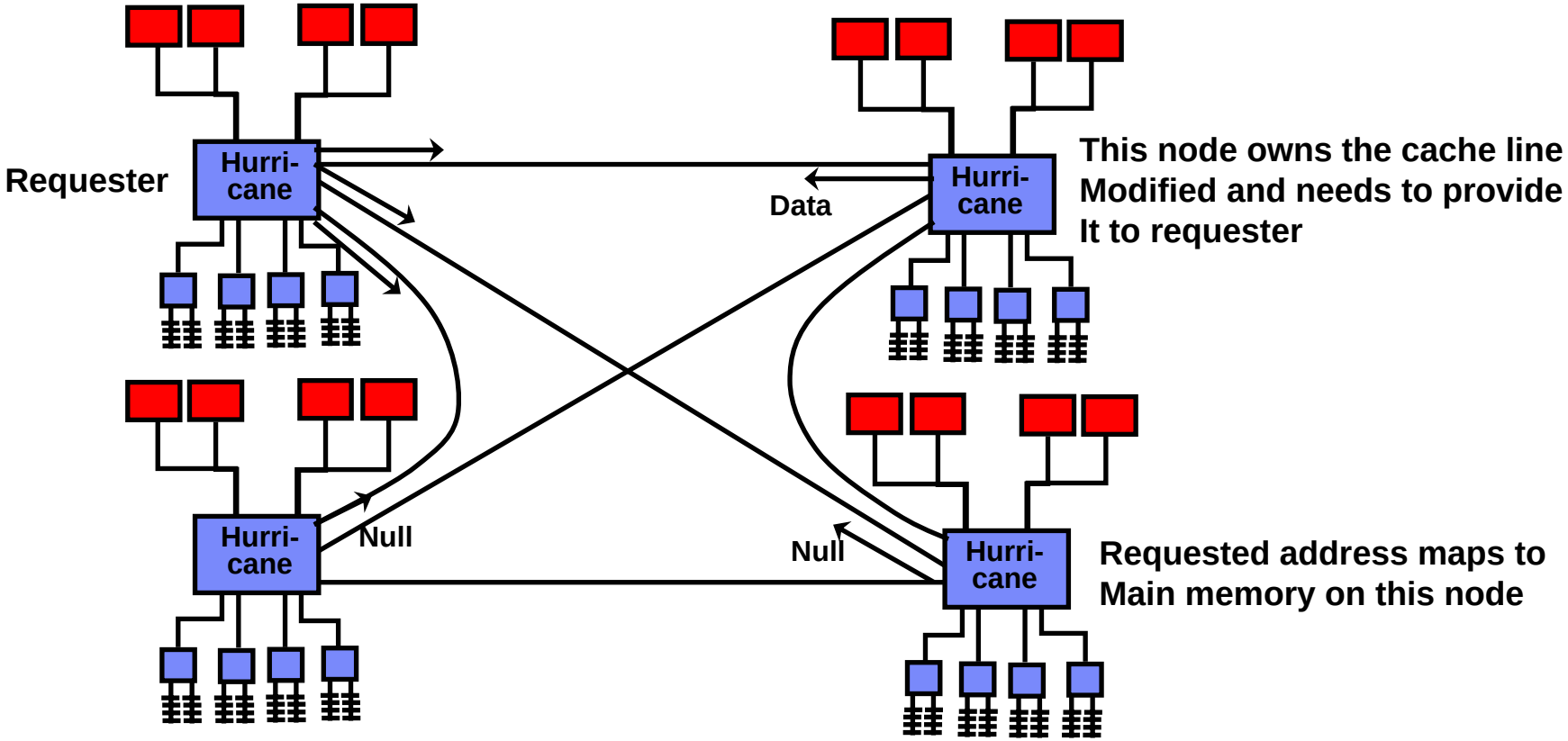




EDRAM Organization for Snoop Filter

* Snoop filter tracks and records all processor cache line requests
* 8 banks of 6Mb EDRAM = 48Mb Total
* Each row is 256 bits wide
* 256 bits=9 way associative with MESI+Tag
* 64B cache lines, 2 way sectored
* Enough Dir entries for 216MB of data
* In Multi-Node system, off node data is backed up in virtual L4 Cache

EDRAM Remote Directory: Multi-Node Example



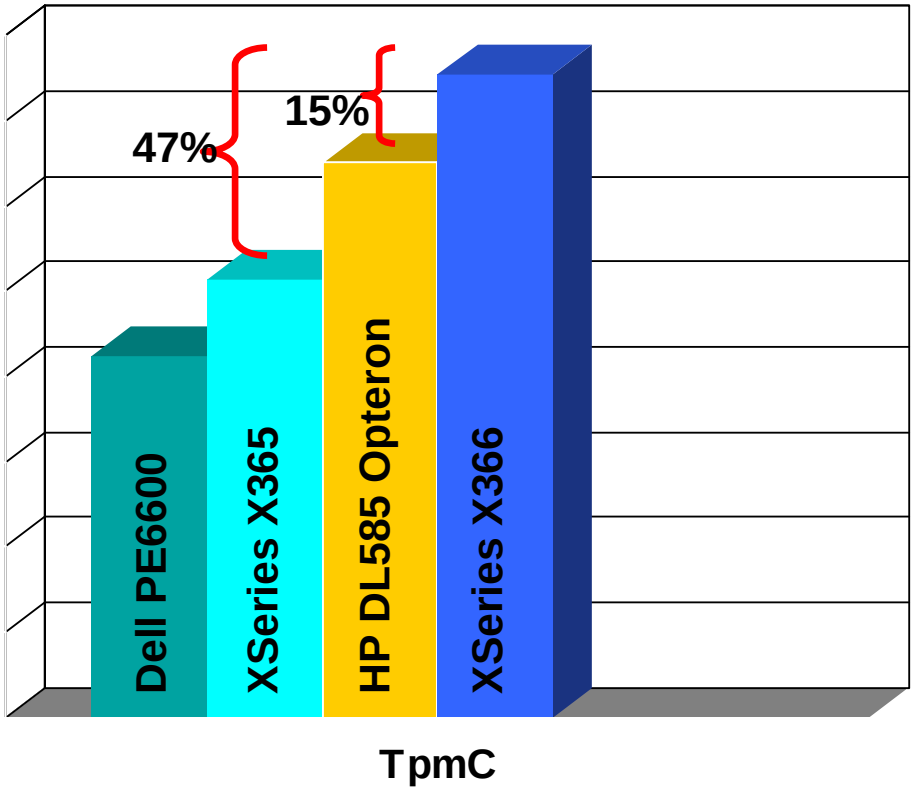
* In Multi-Node system, EDRAM associativity is split into Snoop filter and Remote Directory
* Example configuration: 7-Way Snoop Filter and 2-Way Remote Directory
* Remote Directory tracks local main store data that is checked out by another node
* RDIR required to make EXA multi-node broadcast coherency work – only one node can return data

Recapping Major Changes EXA 2 to X^3

- Single Front Side Bus split in two
  - Doubling connections between processors, memory and I/O
  - Fewer FSB electrical loads allows higher frequency operation
  - Net of 3X FSB bandwidth improvement
- Node control chip (Twister) merged with Memory control chip (Cyclone)
  - Significant increase in main store bandwidth
  - Significant reduction in main store latencies
  - Lower overall costs
- 400MHz DDR L4 cache chip removed
  - Main memory latency reductions obviate need for dedicated L4
  - Virtual L4 technology improves multi-node scalability
- I/O controller chip (Winnipeg) upgraded to PCI-X 2.0 (Calgary)
  - Improved bandwidth and RAS
  - PCI-X 133GHz -> PCI-X 2 266MHz slot bandwidth
- Major focus on remote latency reduction
  - Results in multi-node system scaling like a flat SMP

Breakthrough commercial performance

- IBM eServer X3 + Intel 64-bit Xeon MP:
  - x366: the *fastest* 4-way x86 server on the planet
  - 150.7K tpmC
    - 47% performance increase over x365
    - 15% faster than the previous #1 (Opteron)
  - HP/Dell use Intel's Twin Castle Chipset
    - 31% slower at same cost
    - 25% slower at +$10,916
  - Dell/HP unlikely to publish Twin Castle TPC benchmark



Designing X³

Enterprise X-Architecture™

Historical Timeline

Université des partenaires 2005

1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007

EXA I

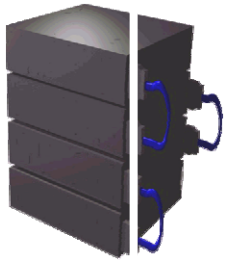
- Pay-as-you-grow scalability with Intel Architecture
- Mainframe inspired RAS

Arch — Design — Validate — GA — Product

EXA II

- Latency reductions; bug fixes
- Laminate packaging
- Dropped future Itanium funding

Arch — Design — Validate — GA — Product



EXA III

- Supports next generation Intel processors
- Significant performance improvements, major redesign
- Cost improvements

Arch — Design — Validate — GA — Product

EXA IV

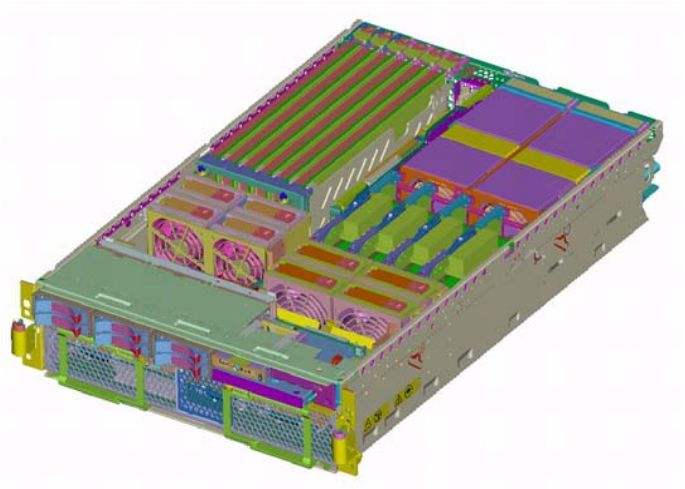
Arch — Design — Val — GA — Prod

- Next Generation Intel Processors
- IO Virtualization

Driven by IBM Hi End Industry Standard Server differentiation

PCI-X 2.0 or PCI-E?

- PCI-X 2.0 meets bandwidth requirements for everything except high speed graphics
  - High speed graphics not critical in servers
  - Fibre Channel and Ethernet vendors support PCI-X 2.0
- PCI-X 2.0 is evolution, not revolution
  - PCI-X 2.0 maintains customer adapter investment
- Decision made to support only PCI-X 2.0 at EXA II architecture closure in 2002
  - EXA III time-to-market deemed more critical than need for PCI-E
  - PCI-E definition did not align with EXA III design schedule
  - No PCI-E south bridges exist - only PCI-X

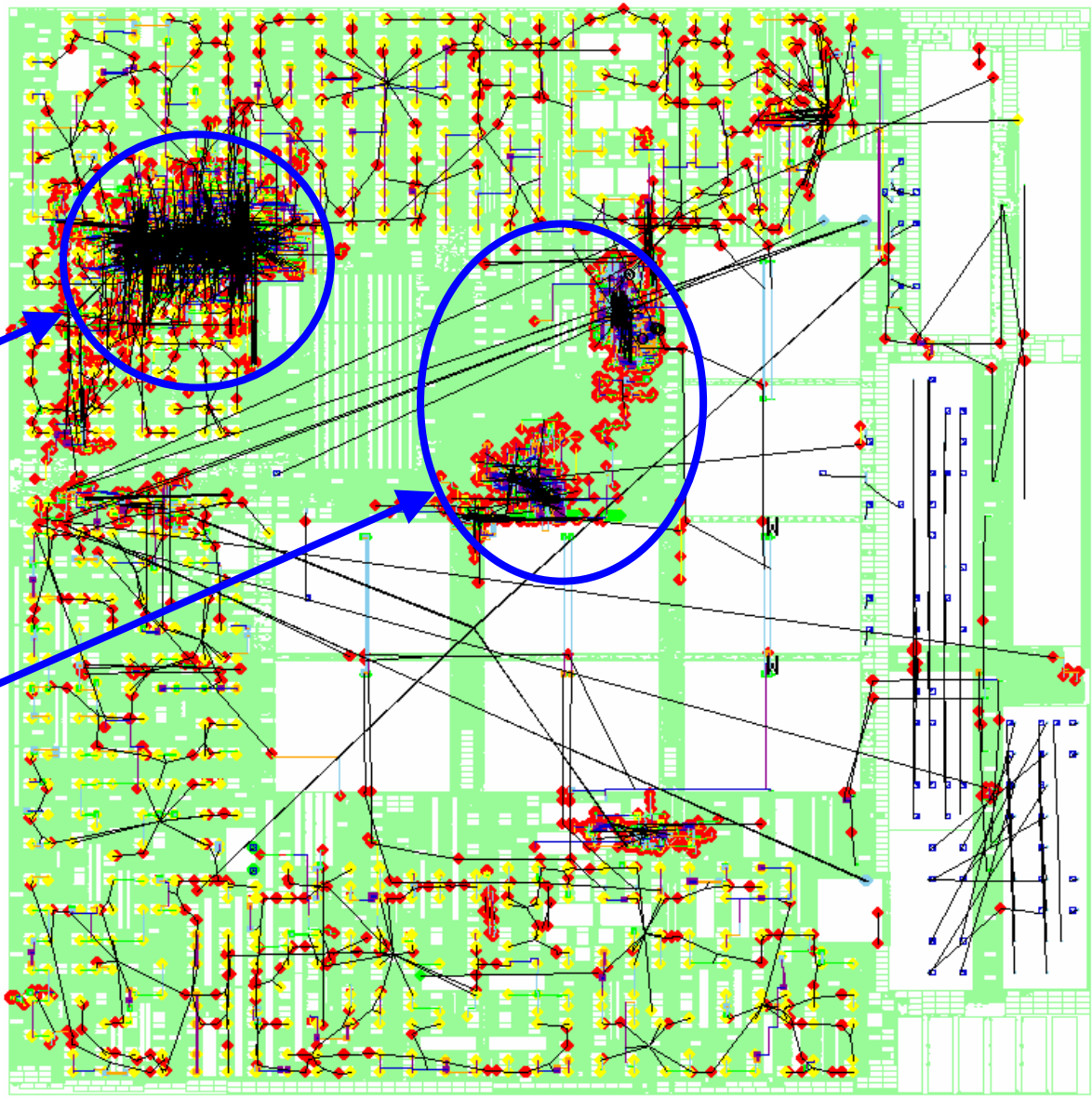


IBM accelerating PCI-E based on Intel dual-processor platform introduction; 1Q06 target to support 50/50 PCI-E/PCI-X 2.0 slots

Designing EXA III – Wiring Challenges

- Biggest challenge on EXA III was wiring Hurricane node controller
  - Wiring problems delayed first tape out by two months
- Die size selection is balancing act
  - IMD ASIC die sizes come in discrete steps
  - Need to keep die small to minimize chip cost
  - Need die large enough to hold all logic
  - Die might be able to hold all logic but still not wire and time
    - Or wiring and timing may be very difficult and take more days
    - Delicate balancing act
- Hurricane die was too small to wire all logic and hold schedule
  - Congestion was in area of transaction sequencers
  - Transaction sequencers reduced from 48 to 32 for first pass
  - Logic redesigned to reduce interconnect for Pass 2
    - Increased transaction sequencers back to original 48

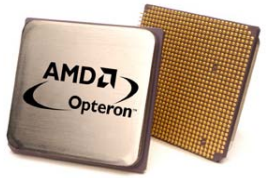
Hurricane Wire Overflows



Designing EXA III – FSB design challenges

- Optimizing performance for both Intel MP and DP based processors
  - MP processor support was base plan
  - EXA history with Prescott/Gallatin indicated value of DP support
- Intel MP Front Side Bus Operation diverges in 1H05
  - DP processors continue to operate in-order (speed clock)
  - MP processors have enhanced deferred phase (L3 cache)
- Decision was made to optimize Hurricane for both
  - Increased schedule risk
- In hind sight, it was the right decision
  - Potomac frequency dropped from 3.6 to 3.3
  - Potomac L3 cache latency increased
  - Hurricane provides better performance with DP processors

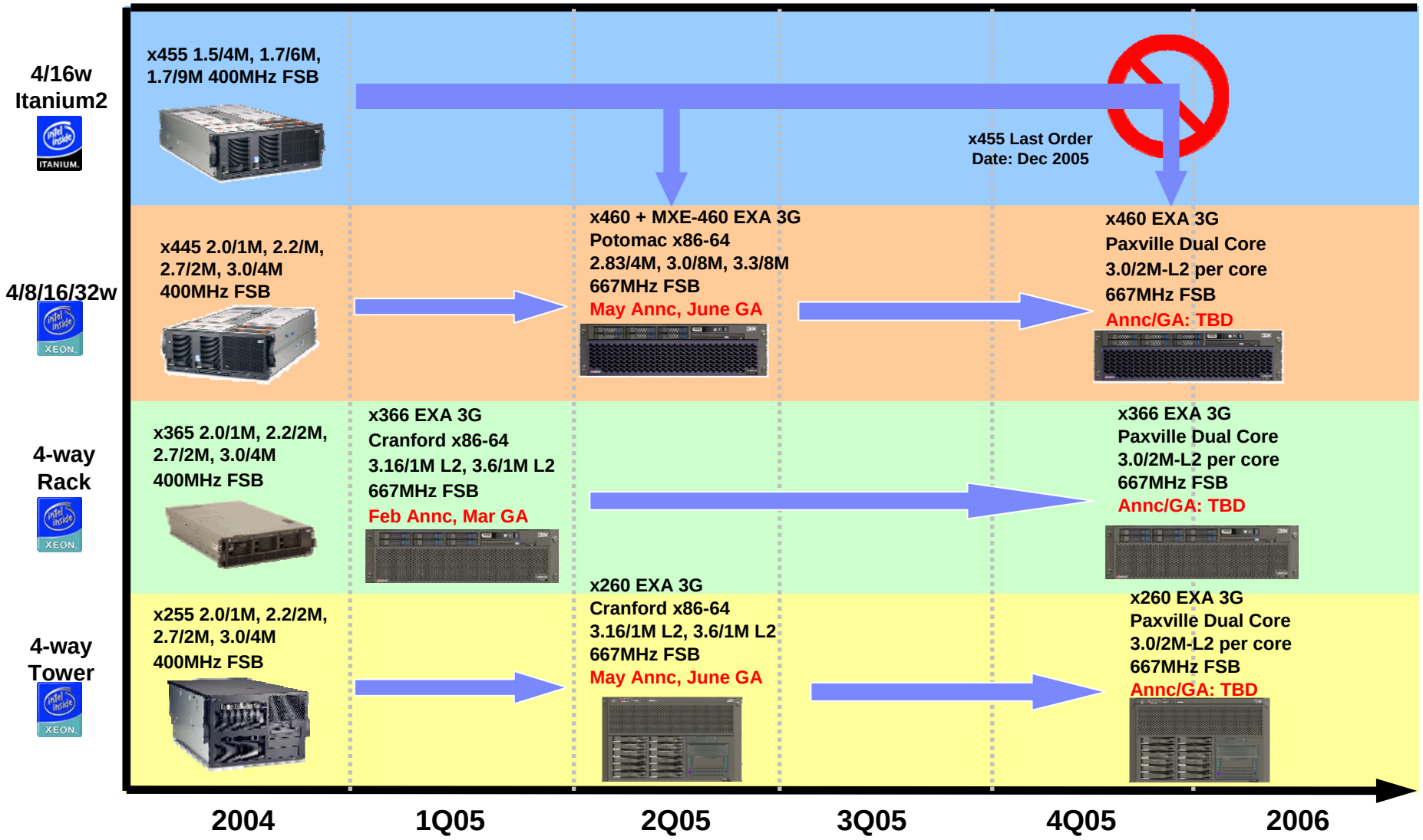
X3 Competition (or lack of it)



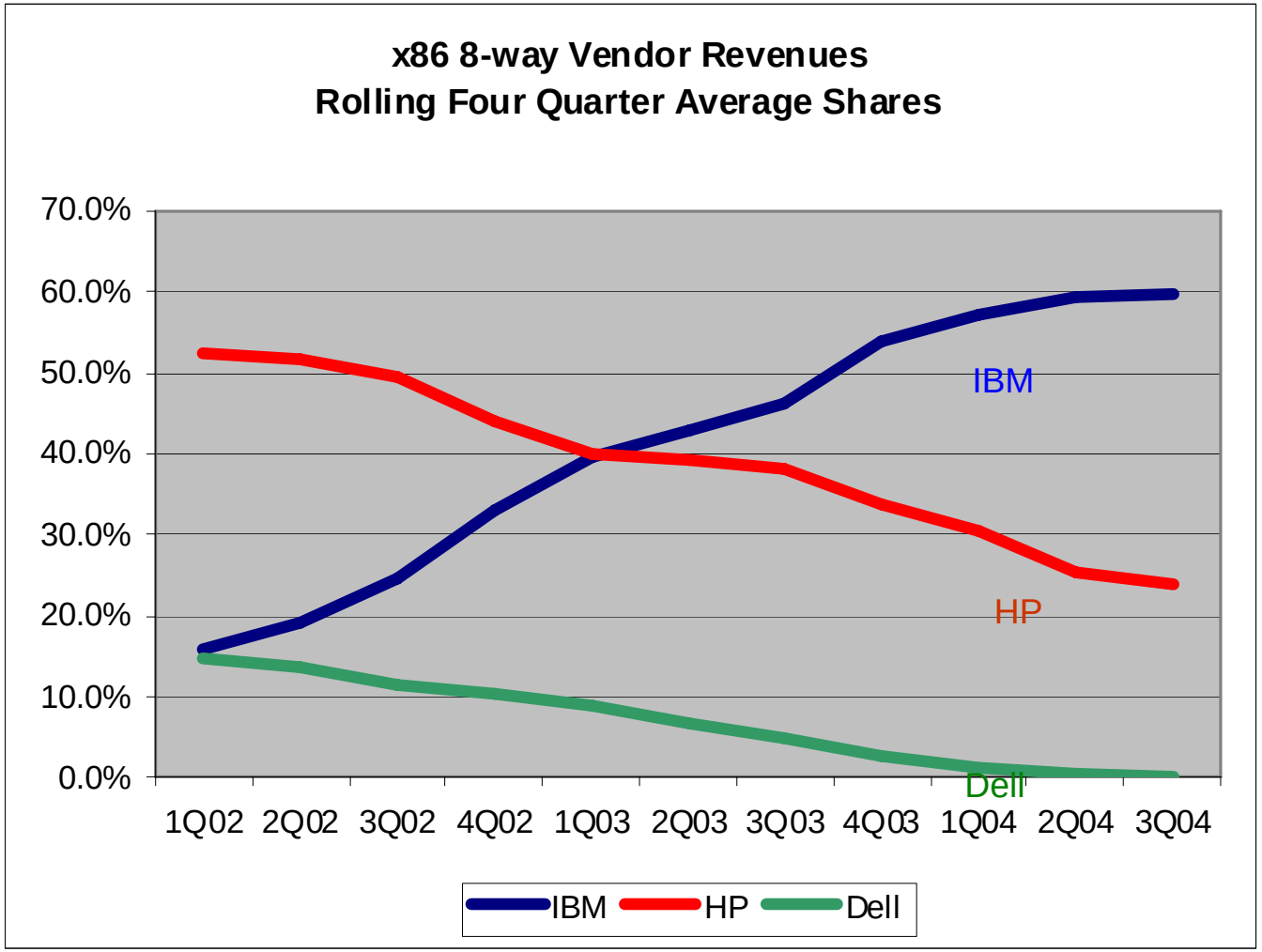
- Intel Twin Castle (Dell, HP, ..)
  - Twin Castle lacks snoop filter
    - All FSB traffic needs to be mirrored to twin bus
    - FSB bandwidth stressed with Cranford, dual-core Paxville makes it worse
    - Greater traffic means higher latency and lower performance
  - X^3 is third generation with strong focus on latency reductions
    - Primary focus on first generation chips was functionality
      - Once defined, performance was focus for subsequent generations
    - Twin Castle is 1st gen with memory latencies up to 2X greater than X^3
  - Twin Castle has no scalability beyond 4-way
  - Twincastle is Dell/HP's only option for Intel based MPs
- AMD Opteron (HP)
  - SMP capped at 4 sockets, no SMP growth path for customer
    - Node controller required, Newisys design dated and not likely to make it to market
  - X3 outperforms Opteron with equal number of cores at all points in time
  - X^3 has significant "Big Iron" RAS features absent on Opteron
    - mirroring, hotplug, RBS, …
  - Not mature enough for conservative Enterprise customer base
    - Jury still out at large companies, hesitancy bolstered by company performance

X$^{3:}$ A Call to Action

xSeries high performance 2005 roadmap

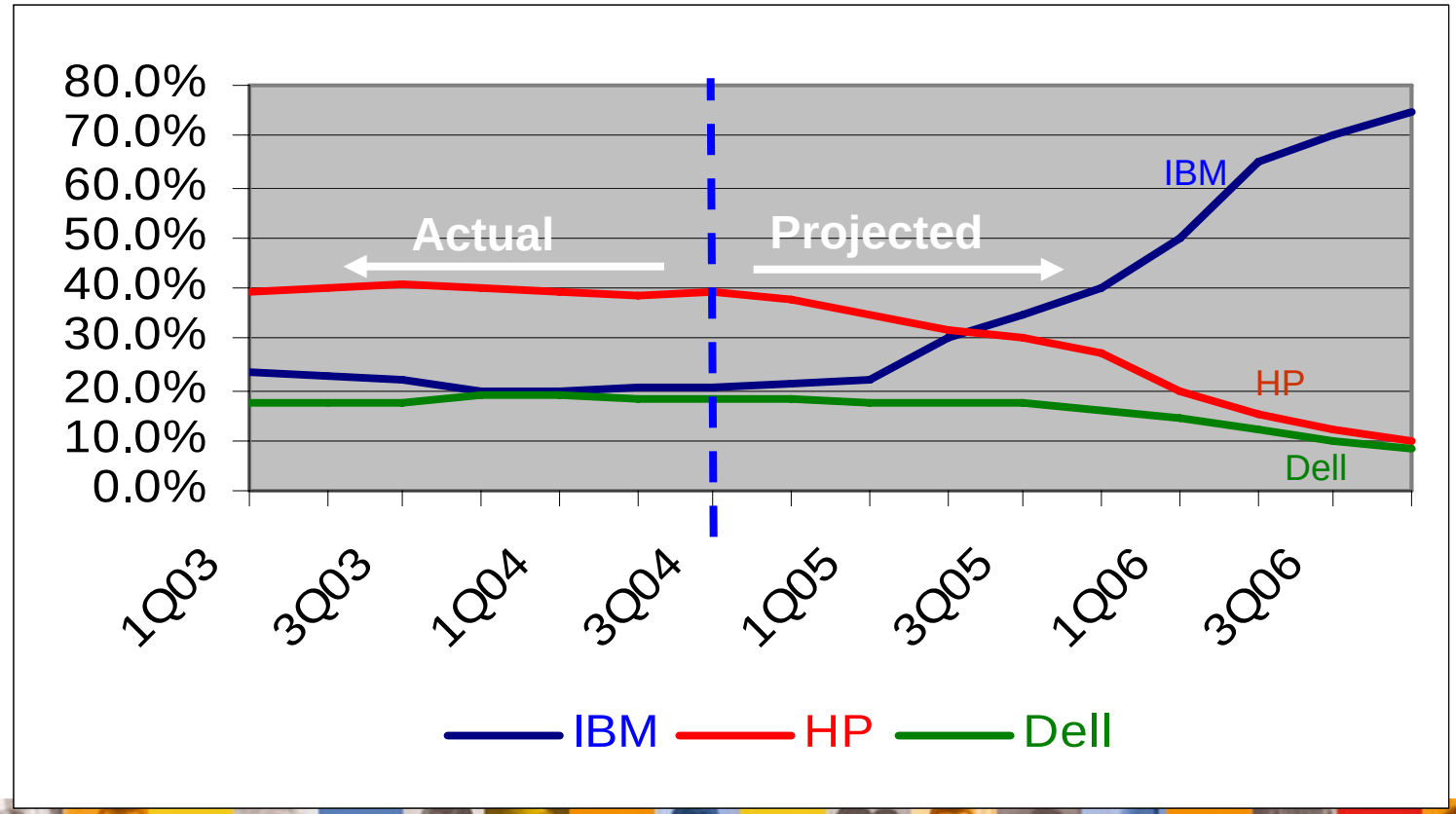


EXA1/2 enabled xSeries to dominate the 8 way market



is the weapon for IBM 4-Way domination

- Neither HP or Dell can compete on performance, cost, or features
- EXA III outperforms Opteron at all points in time with same # cores
- Let's use this new weapon to gain market share in the 4-way space



Trademarks

End of Slides